Topic Sensitive SourceRank: Extending SourceRank for Performing Context-Sensitive Search over Deep Web

by

Manishkumar Jha

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved August 2011 by the
Graduate Supervisory Committee:

Subbarao Kambhampati, Chair
Hasan Davulcu
Huan Liu

ARIZONA STATE UNIVERSITY

December 2011

ABSTRACT

Source selection is one of the foremost challenges for searching deep-web. For a user query, source selection involves selecting a subset of deep-web sources expected to provide relevant answers to the user query. Existing source selection models employ query-similarity based local measures for assessing source quality. These local measures are necessary but not sufficient as they are agnostic to source trustworthiness and result importance, which, given the autonomous and uncurated nature of deep-web, have become indispensible for searching deep-web. SourceRank provides a global measure for assessing source quality based on source trustworthiness and result importance. SourceRank's effectiveness has been evaluated in single-topic deep-web environments. The goal of the thesis is to extend sourcerank to a multi-topic deep-web environment. Topic-sensitive sourcerank is introduced as an effective way of extending sourcerank to a deep-web environment containing a set of representative topics. In topic-sensitive sourcerank, multiple sourcerank vectors are created, each biased towards a representative topic. At query time, using the topic of query keywords, a query-topic sensitive, composite sourcerank vector is computed as a linear combination of these pre-computed biased sourcerank vectors. Extensive experiments on more than a thousand sources in multiple domains show 18-85% improvements in result quality over Google Product Search and other existing methods.

DEDICATION

To my parents.

ACKNOWLEDGEMENTS

I would like to thank my advisor Prof. Subbarao Kambhampati for the opportunity to work under his guidance. I am very thankful for his insightful guidance, help and support.

I would also like to express my gratitude to Prof. Huan Liu and Prof. Hasan Davulcu for their valuable guidance and for being on my thesis committee.

I would like to thank Raju, my colleague at DB-Yochan, for his invaluable inputs and immense help during the course of my research. I am also thankful to Rohit, Sushovan and Yuheng who have been my friends and colleagues at DB-Yochan and have always helped and supported me during my research. I would also like to thank entire planning group at Yochan-lab for their support.

I am thankful to my roommates Adhar, Hardeep and Vishal for their encouragement and for being there during difficult times. I am also thankful to my friends Mahima, Archana, Kuhu, Bernie and Anwasha for their help and support.

I would like to thank Mike and all staff members at Hispanic Research Center for their help and support.

TABLE OF CONTENTS

Page

LIST OF FIGURES vi

CHAPTER

1 INTRODUCTION 1

1.1 Deep Web 1

1.2 Searching over Deep Web 1

1.3 Source Selection 3

1.4 Organization of Thesis 4

2 RELATED WORK 6

3 SOURCERANK 7

3.1 Agreement Computation 7

3.2 SourceRank Computation 8

4 EXTENDING SOURCERANK TO MULTI-TOPIC DEEP-WEB ENVIRONMENT . . 11

4.1 Online Approach - Query-Specific SourceRank, QSR 12

4.2 Offline Approach - Query-Agnostic Undifferentiated-SourceRank, USR . . . 13

5 TOPIC-SENSITIVE SOURCERANK - TSR 17

5.1 Computing Topic-Specific Importance Ranking 19

5.2 Query-Time Processing 19

Training Data 20

Classifier 20

5.3 Source Selection 21

5.4 System Architecture 21

6 EXPERIMENTAL SETUP 23

6.1 Deep-Web Sources S 23

6.2 Sampling Queries Q_S 23

6.3 Test Queries Q_U 24

6.4 Source Selection Models 24

Importance-Based Source Selection Models 24

Query Similarity Based Measures 28

6.5 Result Merging and Ranking 28

Chapter Page
6.6 Relevance Evaluation 28
7 RESULTS 30
7.1 Comparison with Query Similarity Based Source Selection 30
7.2 Comparison with Agreement Based Source Selection 32
7.3 Comparison with Oracular Agreement Based Source selection 34
8 CONCLUSION 35
REFERENCES 36

LIST OF FIGURES

Figure		Page
4.1	Comparison of USR and DSR	15
4.2	Topic-Class Based Comparison of USR and DSR	16
5.1	Multi-Domain Deep Web Integration System Combining Online Query Classification and TSR Based Source Selection.	22
7.1	Comparison of TSR and Relevance-Based Source Selection Models	30
7.2	Topic-Class Based Comparison of TSR and Relevance-Based Source Selection Models	31
7.3	Comparison of TSR and Agreement-Based Source Selection Models	31
7.4	Topic-Class Based Comparison of TSR and Agreement-Based Source Selection Models	32
7.5	Comparison of TSR and Oracular Agreement-Based Source Selection Model	32
7.6	Topic-Class Based Comparison of TSR and Oracular Agreement-Based Source Selection Model	33

Chapter 1

INTRODUCTION

1.1 Deep Web

Traditional web search has been limited to searching over surface-web, the part of web comprising of static html pages. Surface web comprises of only fraction of data available over the web. Recently with many databases being published online, a new type of web content, called deep-web, is available over the web. Deep-web is the collection of web of data stored in databases, concealed behind html query forms. Deep-web data is available in the form of dynamic pages generated in response to query requests made to deep web sources. Traditional search engines rely on the hyper-linked structure and static nature of surface web to crawl and index surface web content. As neither of these are applicable to deep web, search engine crawlers are unable to proceed beyond query forms, failing to extract deep-web content. This makes deep-web literally invisible or hidden to surface-web search engines.

Deep-web sources contain structured data and the information contained in these sources span all the humanly definable topics [7]. Some estimates have pegged the size of deep web to be 500 times that of surface web and the quality of data to be three times that of surface web [7]. Searching over deep-web has become one of the most prominent research areas in information retrieval due to the vast and comprehensive coverage of deep-web content. Unavailability of direct access to deep-web data is one of the major hindrance in searching deep-web as it prevents direct implementation of surface web's information retrieval techniques. Searching over deep web has been identified as the next big challenge in information integration [19].

1.2 Searching over Deep Web

Surfacing and mediator systems have emerged as the two broad categories of search strategies for searching deep-web. Surfacing or data warehousing [13] transforms the dynamic model of deep-web into a static one by precomputing query submissions for all html forms. The idea is to crawl various web-query forms, run queries on each of these forms, collect results and index the resulting page urls and results. Once the contents of

deep web sources have been converted into static pages and indexed, surface-web's information retrieval techniques can seamlessly be used for searching over deep-web. Although this approach provides an innovative way of simultaneously querying over surface and deep web, its main drawback is that it tries to convert the dynamic model of deep web into a static one. When a user actually views these pages as a result of some user-query, the results are stale and may not be accurate. This is specially true for shopping-related deep-web sources as their contents change quite frequently. In addition to reduction in precision, it leads to user dissatisfaction. Another issue with this approach is that it requires blasting deep web sources with unwarranted queries. This can be too taxing for deep web sources.

Mediator or federated information retrieval systems, broker user query simultaneously over a subset of deep web sources, collect the responses and return a ranked set of results to the user. Although mediating is much more difficult than surfacing, it produces more timely and satisfactory results. A naive approach for mediator systems would be to send the user query to *all* deep web sources, collect the results, rank them and present them to the user. This approach is however quite inefficient, too burdening on deep web sources and wastes lot of resources such as network bandwidth and processing power. Majority of deep-web sources may not even be able to answer the query. A better approach is to select the best subset of sources which are expected to provide relevant results for user query. But in order to make informed decision in terms of selecting a subset of sources, mediator systems need information about the content of deep web sources. Over the web there are very few deep web sources which are cooperative and make their entire corpus vocabularly and corpus statistics available to mediator systems. Majority of deep web sources are non-cooperative as access to their content is restricted to query-forms and only provide results in response to submitted queries [7]. Once the query results are returned by selected deep web sources, another challenge for mediator systems is to merge and rank these results irrespective of the ranking provided by deep web sources. This work looks at the source selection problem and tries to address some of its issues.

1.3 Source Selection

Given a user query, the source selection problem is to pick a subset of deep web sources expected to provide relevant results for the query. Even though there has been plenty of research in this area, in both text and relational database community, all efforts have concentrated on evaluating source quality based on query-similarity based relevance measures specifically they estimate the likelihood of a source providing relevant answers for the user query. These source selection methods use local measure for evaluating source quality i.e. source quality is dependent on information that a source provides about itself. Given the uncontrolled and open nature of deep web, another orthogonal but important property to be considered during source selection is that of source trustworthiness and result importance. Over the deep web, there may be hundreds or thousands of sources which are equally relevant to a user query causing an abundance problem. It is important that the mediator system identify trustworthy sources as it is quite possible that some of the relevant sources might have artificially boosted their ranking for economic gain.

Consider a scenario where there are two deep-web sources $amazon.com$ and an untrustworthy source like $xyz.com$. In order to lure users to their site, $xyz.com$ may misrepresent the information by advertising products at deep discounts. Intuitively the mediator system should rank a trustworthy source like $amazon.com$ higher than $xyz.com$. Relevance based measures will fail to identify this and will rank $amazon.com$ and $xyz.com$ equally. When the user clicks on results from $xyz.com$, the user might either be misled on the product price during checkout or might be shown a completely different product severely affecting user satisfaction. This necessitates the need for an additional metric of source trustworthiness for assesing deep-web source quality.

Now that the necessity of having a trustworhiness measure has been established, what should such a measure comprise of? Here are some reasonable desiderata which a deep web source selection model should include,

1. Source selection model should consider source trustworthiness while assessing source quality.

2. Source trustworthiness of a source should be assessed based on a global measure. It should'nt depend on any information that a source provides about itself but on the endorsement of the source by other sources i.e. what the other sources say about a particular source.

The challenges for deep web are similar to those faced by surface web for determining page importance. Computation of page importance in surface web was aided by the existence of hyper-links between web pages. These hyper-links gave rise to an explicit endorsement structure between web pages. Authorities & Hubs [14] and PageRank [8] are some of the earliest and popular surface web algorithmic tools which exploited the linked structure of the web for identifying important or trustworthy pages. Direct application of these surface web techniques is difficult for deep web as no such endorsement structure exists between deep web sources.

At present, SourceRank [6] is the only work which addresses this issue. It introduces an agreement based technique for implicitly creating an endorsement structure between deep web sources. Although sourcerank has been shown to be effective in identifying trustworthy and important sources, its effectiveness has only been evaluated in single-topic deep-web environments. Given the enormous size of deep web, it is difficult to create and maintain such single-topic environments for all topic-classes.

As part of this thesis, automated ways of extending sourcerank to multi-topic deep-web environments are explored. Topic-sensitive sourcerank is presented as an automated, efficient and effective way of capturing source trustworthiness and result importance in a multi-topic deep-web environment.

1.4 Organization of Thesis

Chapter 2 contains a summary of source selection models for deep-web and related work. As sourcerank is central to this work, a brief overview of sourcerank computation is provided in Chapter 3. Chapter 4 discusses the types of deep-web environments and

explores ways of extending sourcerank to a multi-topic deep-web environment. Chapter 5 provides a detailed description of the proposed approach, topic-sensitive sourcerank. Experimental setup is discussed in detail in Chapter 6. Chapter 7 provides results of experiments. Chapter 8 provides a conclusion of the work.

Chapter 2

RELATED WORK

Collection selection has received lot of attention in both relational and text databases community. In addition to relevancy, current relational database selection models use coverage to minimize the cost of retrieving maximum number of unique records from minimum number of sources [15]. Coverage of a database is a measure of number of relevant tuples to the query.

CORI [10] and GlOSS [11] are among the earliest source selection techniques employed for text databases. These techniques are purely query-based relevancy measures. ReDDE [18] considered database size for estimating distribution of relevant documents. Some of the current research [17] has been directed towards considering source coverage and source overlap for minimizing retrieval costs. SourceRank [6] introduces an orthogonal domain-independent global measure for evaluating source quality based on trustworthiness and result importance.

Deep-web sources are non-cooperative as the sources only provide query-based access to their content. Callan and Connell [9] proposed a query-based sampling, QBS, technique for obtaining resource descriptions. In QBS, probe queries are sent to collections and the results returned are used as resource descriptions.

In surface-web, Authorithies & Hubs [14] and PageRank [8] are among the earliest and most popular link-based techniques for identifying important, trustworthy pages. These techniques use the hyper-linked structure of surface web to extract useful information about page importance. Topic-sensitive pagerank [12] presented a topic-based approach for imporving effectiveness of pagerank over surface web.

Chapter 3

SOURCERANK

As described earlier, the absence of explicit endorsement structure between deep web sources is a major hindrance for the application of link-based ranking strategies for deep web. This chapter provides a review SourceRank [6], a measure which evaluates source quality based on trustworthiness and result importance and its computation details.

SourceRank introduces a domain-agnostic agreement-based technique for implicitly creating an endorsement structure between deep web sources. The paper presents and supports the argument that agreement of answer sets returned by deep-web sources in response to same queries, manifests a form of implicit endorsement among deep web sources. This endorsement is modeled as a directed weighted agreement graph where nodes represent deep web sources and edge weights correspond to the agreement between deep web sources. SourceRank, a measure of quality of a source based on trustworthiness and result importance, is computed as the stationary visit probability of a weighted random walk performed on this agreement graph. SourceRank is computed once for each deep-web crawl and all computations are offline. At query-time, a weighted combination of sourcerank and a relevance-based measure is used for ranking deep-web sources based on releavance, trustworthiness and result importance.

3.1 Agreement Computation

Computing agreement among deep-web sources based on answer sets of same query is not trivial. Various sources represent same entity differently rendering equality-based comparisons almost ineffective. Semi-structured nature of deep-web entities provides an interesting middle ground between fully-structured relational database tuples and free-text of text databases. SourceRank work combines and extends record linkage model used in structured relational databases and named entity matching used in free-text IR systems for accurate and timely agreement computation. Agreement is computed using a three-level similarity computation, details of which can be found in the paper [6].

1. Attribute value similarity, $SIM(v_i, v_j)$, estimates the similarity between a pair of attribute-values v_i and v_j, using Soft TF-IDF with Jaro-Winkler as the similarity measure.

2. Tuple similarity, $S(t, t')$, computes similarity between a pair of deep-web entitiy t and t'.

$$S(t, t') = \frac{\sum_{(v_i \in t, v_j \in t') \in M} w_{ij} SIM(v_i, v_j)}{\sqrt{\sum_{v_i, v_j \in M} w_{ij}^2}} \tag{3.1}$$

   v_i and v_j are attribute values of tuple t and t' respectively, w_{ij} is the weight assigned to the match between v_i and v_j and M is the matched pairs of attribute values between t and t'. w_{ij} is computed as the mean inverse document frequency of tokens in v_i and v_j.

3. Result set agreement, $A(R_{1q}, R_{2q})$, computes the agreement between the result sets R_{1q} and R_{2q} returned by deep web sources S_1 and S_2 respectively in response to query q.

$$A(R_{1q}, R_{2q}) = \sum_{(t \in R_{1q}, t' \in R_{2q}) \in M} S(t, t') \tag{3.2}$$

   M is the matched pairs of tuples between result sets R_{1q} and R_{2q}.

Overall agreement between a pair of sources S_1 and S_2, A_{Q_S}, is the aggregate of agreements for sampling queries Q_S.

$$A_{Q_S}(S_1, S_2) = \sum_{q \in Q_S} \frac{A(R_{1q}, R_{2q})}{|R_{2q}|} \tag{3.3}$$

SourceRank employs a greedy technique for pair-matching operations. This helps restrict the agreement computation time to $O(N_S^2)$ where N_S is the number of deep web sources.

3.2 SourceRank Computation

Agreements between sources are modeled as a directed weighted agreement graph. Graph nodes represent the deep-web sources and edge weights represent the agreement between the sources. To account for sampling bias, smoothing links are added between

every pair of nodes. The weight of an edge $S_1 \rightarrow S_2$ is computed as,

$$w(S_1 \rightarrow S_2) = \beta + (1 - \beta)\frac{A_{Q_S}(S_1, S_2)}{|Q_S|} \tag{3.4}$$

The weight of out-going edges of each vertex is normalized such that they sum to one for each vertex.

A random deep-web searcher model is used for determining source quality based on the agreement graph. According to this model, a searcher who has been provided with the agreement graph will start his search by randomly picking a deep-web source. If the searcher finds the contents of the source useful, then it is highly likely that he will also find useful the contents of sources agreeing with current source. The searcher can select any one of these agreeing sources by selecting one of the outgoing links with a probability equal to the weight of outgoing link and continue his search with the source at other end of the link. If the searcher does not find the content of source interesting, then he can randomly select any source by following one of the smoothing links. Quality of a source is the probability with which a random deep-web searcher will visit the source which is computed as the stationary visit probability of a random walk performed on the agreement graph.

Sourcerank computation can also be explained in terms of eigen vector calculation. Let M be the square stohastic agreement matrix. Value of M_{ij} is the normalized weight of edge $S_j \rightarrow S_i$. Let SR be the sourcerank vector. Initially all sources are assumed to be of same quality. If there are N_S number of sources, then SR is a column vector of size N_S and is initially initialized to $1/N_S$. SourceRank is computed iteratively by multiplyig SR with matrix M and updating SR after each iteration. The iteration stops when SR vector remains unchanged in successive iterations or the change is within threshold, giving the actual sourcerank vector SR^*. SR^* denotes the stationary visit probability of all sources of a weighted random walk performed on the agreement graph.

$$SR^* = M \times SR^* \tag{3.5}$$

To guarantee convergence, M must be irreducible, i.e. the agreement graph should be strongly connected. Smoothing links ensure this and SourceRank computation

converges to a fix point value. Let M' be a matrix with smoothing links, M' can be expressed as,

$$M' = (1 - \gamma) \times M + \gamma \times U \tag{3.6}$$

where γ is the smoothing factor or the weight given to smoothing links and U is the reset distribution matrix representing the smoothing links i.e. $U = [1/N_S]_{N_S * N_S}$

Hence,

$$SR^* = M' \times SR^* \tag{3.7}$$

Chapter 4

EXTENDING SOURCERANK TO MULTI-TOPIC DEEP-WEB ENVIRONMENT

In this section a distinction is made between single-topic and multi-topic deep-web environments and automated ways of implementing sourcerank in these environments are explored.

A vertical mediator system for a topic class c, is essentially a federated information retrieval system, FIR_c developed for a vertical deep-web environment, DW_c, for topic class c. DW_c comprises of a subset of deep-web sources S_c such that sources in S_c contain information related to topic class c. FIR_c expects that the information need of user queries Q_{U_c} posted to such a system are also related to the same topic class c. For FIR_c to be effective, it is essential that the resource descriptions of S_c reflect the coverage of sources S_c with respect to topic class c. SourceRank SR_c created for FIR_c should capture relative trustworthiness of sources in S_c and result importance with respect to topic class c. Both, effective topic-based resource descriptions and sourcerank SR_c can be achieved by using sampling queries Q_{S_c} which are representative of topic class c.

Thus the deep-web environment DW_c of mediator system FIR_c for topic-class c can be defined in terms of sources S_c, sampling queries Q_{S_c} and user queries Q_{U_c} such that $S_c, Q_{S_c}, Q_{U_c} \in c$.

$$DW_c : S_c, Q_S, Q_U \in c \tag{4.1}$$

where c is a topic-class.

Arifare comparison portals are one such examples of vertical applications. SourceRank in combination with a query-relevance based measure has been found to be quite effective in vertical deep-environments like DW_c for topic-class c. The drawback of vertical deep-web environments like DW_c is that they are difficult to create and maintain and address the information need of a very small set of users, those interested in topic class c. Bergman [7] has shown that deep-web contains information spanning all humanly definable topic classes. The enormous size of deep-web makes it extremely difficult if not impossible for creating vertical mediator systems for all topic-classes.

Manually created online, public open-directories like $dmoz.org$ [2], $YahooDirectory$ [5] give a sense of the topic-classes spanned by deep-web. With this information the definition of vertical deep-web environments can be extended to define the components of a multi-topic deep-web environment, DW_C, for a set of topic-classes C,

$$DW_C : S, Q_S, Q_U \in C \tag{4.2}$$

If C^* is the set of topics defined in open-directories, then the complete deep-web environment can be defined as,

$$DW^*_{C^*} : S, Q_S, Q_U \in C^* \tag{4.3}$$

As the deep-web grows, it is highly likely that mediator sytems will have to handle environments like DW_C where the environment contains information related to a set of topic-classes C (multi-topic environments), than DW_c where environment contains information related to a single topic-class c (single-topic environments). It would be interesting to evaluate the effectiveness of sourcerank in multi-topic deep-web environments.

In this section different approaches for extending a source trustworthiness and result importance measure, SourceRank, to a multi-topic deep web are explored. Similar to surface web, there can be two main approaches for extending sourcerank to deep web - an online approach that considers query-time information for identifying trustworthy sources and an offline-approach which is query agnostic. A detailed description of these approaches is provided next along with their evaluation in terms of their feasibility and effectiveness.

4.1 Online Approach - Query-Specific SourceRank, QSR

One approach for extending SourceRank to deep web is to make SourceRank computation query specific i.e. use query-time information for identifying trustworthy sources. This way only the sources relevant to the query topic are ranked. This is similar to HITS [14] used in surface web for identifying authorities and hubs. HITS performs query-time processing on a subgraph of link structure of web to deduce authorities and

hubs. For deep web, the question would be which subset of deep web sources to query in order to compute QSR? As the user is interested only in relevant and trustworthy sources, query-relevance based similarity measures can be used to identify the subset S' of query-relevant deep web sources to be used for computing QSR. With the enormous size of deep web, it is quite possible that size of S' could well be in hundreds of thousands. Polynomial computation time of sourcerank makes sourcerank computation for hundreds of thousands of source infeasible during query-time. For efficiency reasons, computations can be restricted to just *top-k* relevant sources. But it turns out that picking this k is not trivial. HITS provides a desiderata that the sources belonging to S' must satisfy

1. S' should be relatively small
2. S' should be rich in relevant sources
3. S' should contain most (or many) of the most trustworthy sources

In the outlined approach, it is clear that 2. is easily satisfied and 1. has an impact on 3. Having a small value of k will affect 3. HITS found that 3. is typically satisfied when $|S'|$ is between 1000-5000. Using this information, SourceRank can be computed using atleast *top-1000* relevant sources and the *top-k* sources ranked using SourceRank will be relevant and trustworthy. Although this approach will be able to identify important sources for each query, it has its own share of drawbacks. Given the time required for querying sources, retrieving results, computing pair-wise agreement between the sources and computing sourcerank during query-time, this approach is highly inefficient. Also since sourcerank is computed on a very small subset of deep web sources, the approach is susceptible to localized spam. It can also lead to less diversity in the results as the top sources are likely to agree with each other and produce similar results, a problem quite evident in sourcerank for vertical applications.

4.2 Offline Approach - Query-Agnostic Undifferentiated-SourceRank, USR

The other approach for extending SourceRank to deep web is to compute it offline. Although sourcerank will capture trustworthy sources across deep web, it will be query

agnostic. At query-time, a weighted combination of sourcerank and relevance source ranking returned by a query-similarity based measure can be used to get a ranking of sources based on relevancy and trustworthiness. This is similar to the surface web's PageRank [8] approach. In surface web creation of single importance-based ranking is easier as the explicit endorsement structure between all web pages is easily available. In case of deep-web, this endorsement structure is implicitly created using a set of sampling queries. As sourcerank is computed offline, the approach is not only feasible but also efficient in terms of query-time processing.

The drawback with this approach is that the single undifferentiated importance ranking of deep-web sources fails to capture the fact that sources considered trustworthy for some topic-classes may not be considered trustworthy for other topics. For example, consider a deep web environment consisting of books and camera sources. A book source like $barnesandnoble.com$ would be considered trustworthy for queries related to books but highly untrustworthy for camera related queries and the converse is true for a camera source like $jr.com$. This drawback will be quite evident when the deep web environment is dominated by sources containing information related to a subset of representative topics. In this case, sources belonging to the dominating topics will have relatively high sourcerank as compared to other topic sources. This is based on the fact that sources belonging to dominating topics will be quite heavily linked, leading to higher sourcerank for these sources. Thus a single importance ranking approach will not be as effective as a FIR system comprising of mutliple vertical systems, one for each of the representative topics.

This was also evident from the experiments carried out on a four-topic deep-web environment. For evaluation, the performance of FIR system, using USR computed for the four-topic deep-web environment, was compared with that of a FIR system comprising of four vertical systems, DSR, one for each of the topic-class. Each of these domain-specific vertical systems DSR_i where $i \in C$ used a relevance measure and sourcerank for computing source ranking. The purpose of comparing USR with a system like DSR comprising of vertical systems is that DSR will provide an upper bound on the optimum precision that can be achieved by combining relevance measure with source

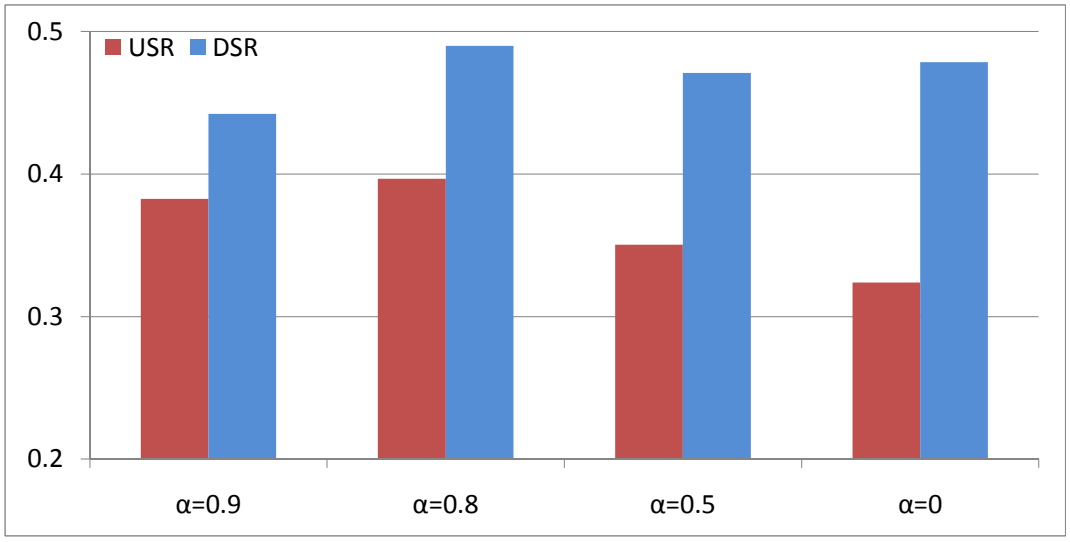

Figure 4.1: Comparison of USR and DSR

trustworthiness. Experimental setup section provides detailed information about the two setups. Using test queries which were a mix of the four-topic classes, the precision of the two systems - USR and DSR, was computed. Based on the experiments, there was $20\% - 50\%$ difference in precision values between USR and DSR as shown in Figure 4.1.

After performing a per topic-class analysis of test queries for $alpha$=0.9, it was found that USR was able to match DSR performance in one topic-class and there was a significant drop in its performance for the remaining topic-classes. As shown in Figure 4.2, USR is not able to identify important sources for $Camera$, $Movie$ and $Music$ topics as effectively as DSR, which is reflected in the drop in precision values for test-queries for these two topics.

Query-agnostic undifferentiated-sourcerank approach turns out to be feasible and efficient for deep-web. But its inability to identify topic-specific importance of sources leads to non-uniform performance across different topics making it less effective.

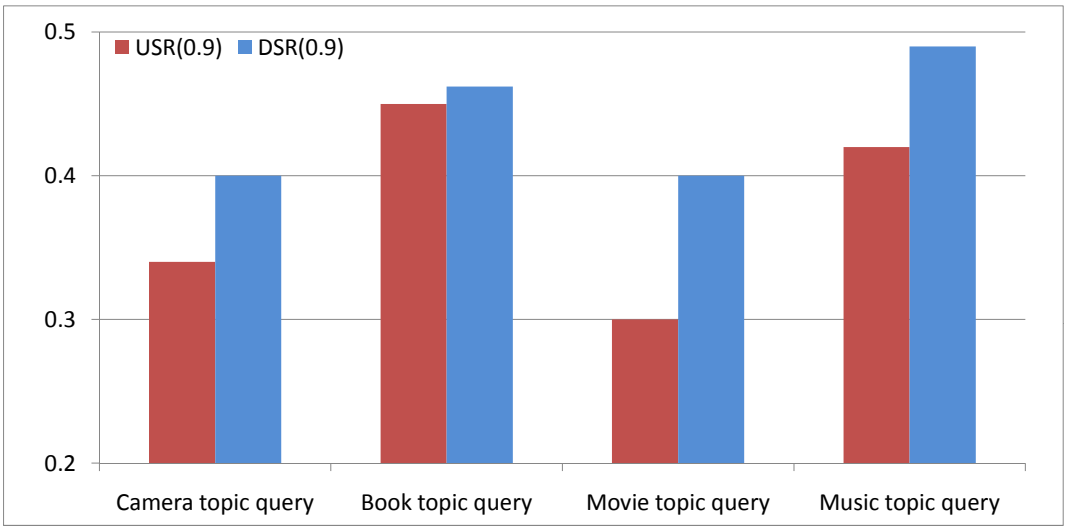

Figure 4.2: Topic-Class Based Comparison of USR and DSR

Chapter 5

TOPIC-SENSITIVE SOURCERANK - TSR

The two most popular surface web link-based ranking strategies for identifying important pages turn out to be either infeasible or ineffective for deep web. An implementation based on HITS approach will impact query-time efficiency, fail to identify all the trustworthy sources, will be susceptible to local spam and will impact the diversity of source selection model. Implementation based on PageRank approach will fail to utilize query-time information for accurate computation of trustworthy sources. Given the unique challenges posed by deep web and polynomial computation time for SourceRank, an ideal approach for effectively extending trustworthiness measure to deep web would be one which is

1. Feasible
2. Requires minimal query-time processing
3. Utilizes query-time information for estimating trustworthy sources with respect to a user query

Based on the earlier definitions for single-topic and multi-topic deep-web environments, these two environments can be viewed as two extremes of the same spectrum. As stated earlier, a deep web environment can be defined in terms of amount of information available with respect to topic association of deep web sources, sampling queries and user queries. A single-topic deep web environment is based on the availability of oracular information that the information contained in deep web sources, sampling and user queries all belong to same single broad-topic. While for a multi-topic environment, the deep web sources, sampling and user queries belong to a broad range of topics with no topic-based classification available for any of these variables. Sourcerank is quite effective in single-topic environments but such environments are hard to achieve and maintain for every broad-topic.

Multi-topic deep-web environments represent the real web scenario, but as seen earlier, sourcerank loses its effectiveness in this environment. In general, agreement by

sources in the same topic-class is likely to be much more indicative of importance of a source than endorsement by out of domain sources. Moreover, sources might have data corresponding to multiple topics. The importance of the source might vary across those topics. For example, Barnes & Noble might be quite good as a book source but might not be as good as a movie source (even though it has information about both topics). These problems are noted for surface web (e.g. Haveliwala [12]), but is more critical for the deep web since sources are even more likely to cross topics/domains than single web pages. To account for this fact, in this work, the deep web source selection is extended by assessing a topic-sensitive quality metric for the sources.

Instead of creating a single importance ranking for all deep web sources, multiple importance ranking of deep web sources are created, each biased towards one of the representative-topic of deep-web environment. Each of these topic-specific importance rankings are computed offline and will capture the relative authority of deep web sources for every topic. At query-time, query-topic is computed i.e. the likelihood of the query belonging to each of the representative topics. Using the query-topic, the individual topic-specific importance rankings are combined to get a query-topic sensitive, composite importance ranking. A conjunction of composite importance scores and relevance scores returned by a query-similarity based measure is used for ranking the sources.

Not only is this approach feasible and efficient in terms of query-time processing, but is also effective as it makes use of query-time information to accurately capture the notion of source trustworthiness and importance for a given query-topic. TSR is not susceptible to localized spam as the individual topic-specific sourceranks are computed on all sources. While computing composite sourcerank, no single precomputed sourcerank vector biased towards a particular topic is picked, instead the individual biased sourcerank vectors are linearly combined based on the fractional topic membership of the query. Avoiding picking winners ensures that diverse sources are selected, ensuring diversity in results. Next section describes the steps for implementing TSR.

5.1 Computing Topic-Specific Importance Ranking

In order to compute biased importance ranking, representative topics for deep-web environment need to be identified. As mentioned earlier, open directories are one of the best source for selecting representative topics for deep web. Open directories are freely available and since they are manually constructed, they closely represent the human notion of topics and topic hierarchies. Topic-sensitive pagerank used 16 top-level categories listed on $dmoz.org$ as a set of representative topics. Query-logs are another way of identifying the broad topics that search engine users are most likely interested in. As deep-web sources are non-cooperative, query-based sampling techniques are used for computing pair-wise source agreement. Each topic under ODP contains links to surface-web sites which are authoritative sources on these topics. ODP along with the web-links to authoritative sources serve as a source for sampling queries to be used for each representative topic.

Using the set of sampling queries for each representative topic and sourcerank agreement computation, biased agreement graphs, AG_c, are computed for each representative topic c, as described in [6] . To account for sampling bias, smoothing links are added between every pair of sources in the biased agreement graphs. Performing a random walk on the biased agreement graphs, AG_c produces topic-specific sourcerank vectors, TSR_c.

5.2 Query-Time Processing

The next set of computations are performed at query time. The first task is to identify the query-topic i.e. the likelihood of the query belonging to representative topic-classes. This can be treated as a soft-classification problem. For a user query q and a set of representative topic-classes c_i where $i \in C$, the goal is to find the fractional topic membership of query q with each of the topics in c_i. For this task, a classifier and training data are required.

Training Data

In order to accurately identify query-topic, training data should be a description of the representative topic-classes. This can only be obtained from deep-web sources by posing the right kind of questions to them. For obtaining topic-descriptions, the questions have to be related to keywords which are representative of the topics. Query-based sampling techniques are used for obtaining topic-descriptions. As the topic-specific sampling queries Q_{S_i} where $i \in C$ are representative keywords of topic-classes, the answers returned by deep-web sources as responses to these sampling queries will contribute towards topic-descriptions. *Top-k* results returned by every deep-web source for topic-specific sampling queries, contribute towards topic-specific descriptions. Answer-set of topic-specific sampling queries are grouped into text documents resulting in a text document D_i for each topic-class c_i. As topic-descriptions are treated as bag of words, topic-statistics for topic-class c_i are nothing but the number of occurrence of terms t in document D_i.

Classifier

The classifier tries to identify the query-topic using query-terms and training data, consisting of topic-class descriptions. In the proposed implementation, a multinomial naïve-Bayes classifier, NBC, is used with parameters set to maximum likelihood estimates to determine the topic probabilities for the user query. When a user submits a query q, the fractional membership of the query is computed for different topic-classes i.e. the topic-class probabilities conditioned on the query q are estimated.

For a topic-class c_i, this is computed as,

$$P(c_i|q) = \frac{P(q|c_i) \times P(c_i)}{P(q)} \propto P(c_i) \prod_j P(q_j|c_i) \tag{5.1}$$

where q_j is the j^{th} term of user query q.

$P(c_i)$ can be set based on availability of domain knowledge but for this work uniform probabilities are used for topic-classes. So the above equation can be updated as,

$$P(c_i|q) = \prod_j P(q_j|c_i) \tag{5.2}$$

After computing the topic probabilities of the query, the next step is to compute the query-topic sensitive importance scores for all deep web sources. For a source s_k, its query-topic sensitive score or the composite sourcerank score, CSR_k is given by,

$$CSR_k = \sum_i P(c_i|q) \times TSR_{ki} \tag{5.3}$$

where TSR_{ki} is the topic-sensitive sourcerank score of source s_k for topic-class c_i

CSR vector gives the query-topic sensitive sourcerank for all deep-web sources.

Since CSR is computed during query-time, it is important that its processing time is kept to a minimal. As CSR will be used in conjunction with a relevance measure, instead of computing CSR for all sources, it can be restricted to just the relevant sources. As long as number of representative topics is small, topic-sensitive sourcerank approach is efficient. For large number of representative topics, CSR computation can be performed by selecting only *top-k* most relevant topics for user query q to minimize query-processing time.

5.3 Source Selection

Source selection for user query q based on relevancy and importance, involves a weighted combination of relevance scores, R returned by a query-based similarity relevance model and CSR computed using TSR approach. For a source s_k, its overall score based on relevancy and importance is computed as,

$$OverallScore_k = \alpha \times R_k + (1 - \alpha) \times CSR_k \tag{5.4}$$

where α is the weight given to query-relevancy model. In this work, α value is experimentally estimated.

Top-k sources for user query q are selected based $OverallScore$ and q is brokered over the selected sources.

5.4 System Architecture

Figure 5.1 provides an overview of the system performing topic-sensitive source selection. It consists of two main parts. An offline component which uses the crawled data for computing topic-sensitive SourceRanks and topic-descriptions. As mentioned earlier, both

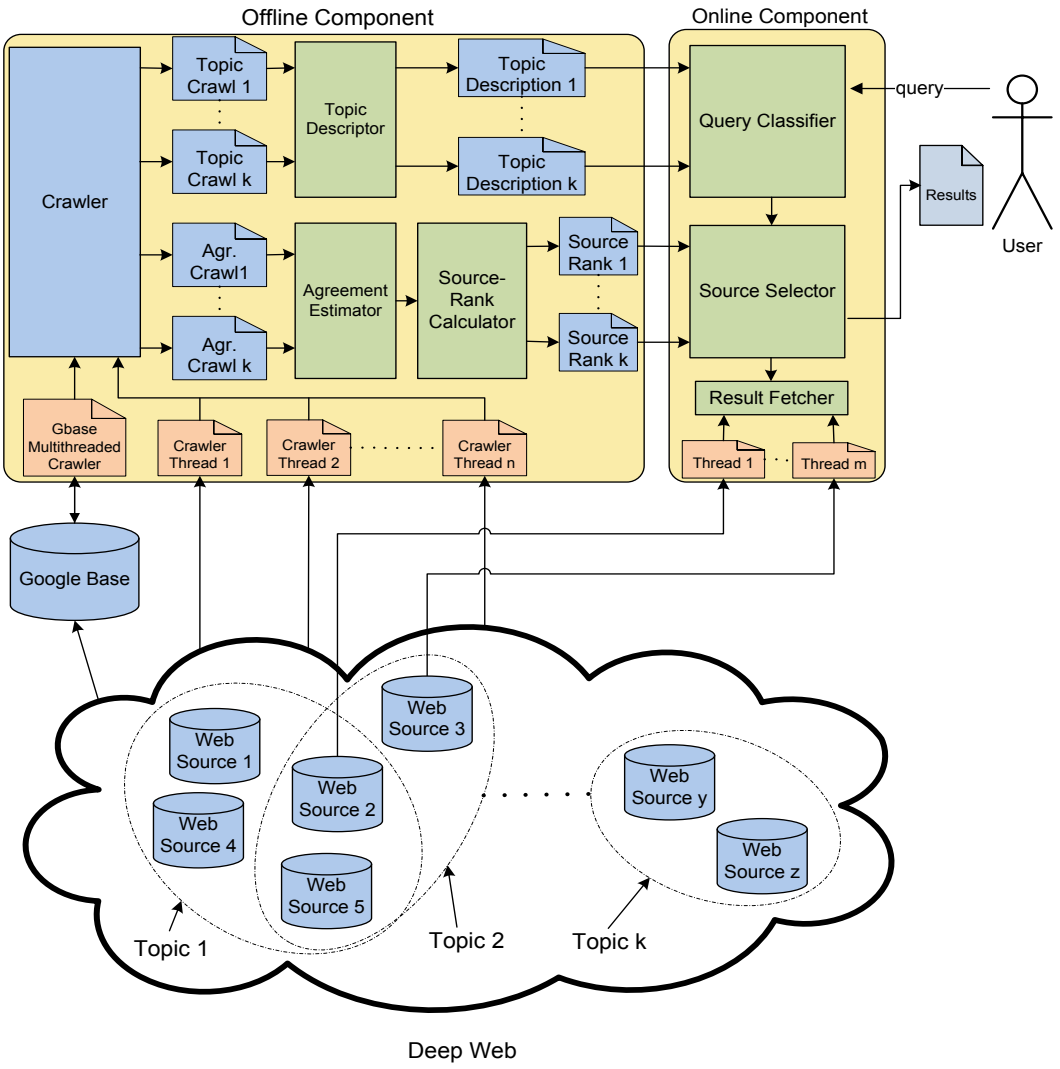

Figure 5.1: Multi-Domain Deep Web Integration System Combining Online Query Classification and TSR Based Source Selection.

these computations are influenced by the topic-specific crawl obtained using topic-specific sampling queries. The online component consists of a classifier which performs user query-classification using the topic-descriptions. The source selector uses the query-classification information to combine TSRs in order to generate query specific ranking of sources. Result fetcher queries the top-k ranked sources, merges and ranks the results and returns top-5 results to the user.

Chapter 6

EXPERIMENTAL SETUP

To evaluate the effectiveness of the proposed approach and compare it with other source selection methods, experiments were performed on a multi-topic deep-web environment with more than thousand sources in four representative topic classes - camera, book, movie and music.

$$C = \{camera, book, movie, music\} \tag{6.1}$$

Equation 6.1 represents the set of topic-class C used for the experiemental setup.

6.1 Deep-Web Sources S

For the experiments, deep-web sources were collected via Google Base. Google Base acts as a central repository where merchants can upload their databases thereby publishing the databases over the web. Google Base provides API-based access to data, returning ranked results. Google Base's Search API for shopping allows querying of data uploaded to Google Base. Each deep-web source in Google Base is associated with a sourceId. For selecting sources for the multi-topic deep-web environment, Google Base was probed with a set of 40 queries. These 40 queries contained a mix of camera model names, book, movie and music album titles. From the first 200 results of each query, sourceIds were collected and these sourceIds were considered as a source belonging to the multi-topic deep web environment. A total of 1440 deep web sources were collected for the multi-topic environment.

$$S = \bigcup_{i \in C} S_i \tag{6.2}$$

Equation 6.2 represents the set of sources S used for the multi-topic environment.

6.2 Sampling Queries Q_S

For the experiments the deep-web sources were assumed to be non-cooperative. Query-based sampling strategy is used for obtaining a sampled set for the sources. For generating the sampling queries publicly available online resources were used. 200 camera model names were randomly selected from pbase.com [3], 200 book titles from New York Times best sellers books [1], 200 movie titles from dmoz.org [2] and 200 music

album names from wikipedia's top-100 number one singles titles from 1986-2010 [4]. A total of 800 sampling queries were used.

$$Q_S = \bigcup_{i \in C} Q_{S_i} \tag{6.3}$$

Equation 6.3 represents the set of sampling queries Q_S used for the multi-topic environment.

6.3 Test Queries Q_U

Test query set contained a mix of queries from all four topic-classes and represents the possible user queries Q_U. Test queries were selected such that there is no overlap with the sampling queries Q_S. The test queries were generated by randomly removing words from camera model names, book, movie and music album titles with probability 0.5. A total of 200 test queries containing 50 queries from each of the representative topic were used.

$$Q_U = \bigcup_{i \in C} Q_{U_i} \tag{6.4}$$

Equation 6.4 represents the set of test queries Q_U used for the multi-topic environment.

6.4 Source Selection Models

This section discusses the experimental setup of different source selection models. TSR is compared with importance based and query similarity based source selection methods. The agreement based methods consider the source agreement, and hence the trustworthiness and relevance of the sources are taken into account. On the other hand, pure query similarity measures like CORI [10] assesses the source quality based on similarity of content with the user query; hence agnostic to the trust and importance.

Importance-Based Source Selection Models

As there has not been any related work on using importance-based measure like sourcerank in multi-topic deep web environment, three mediator systems which employ importance-based source selection models are created and their performance is evaluated based on result precision. As the mediator systems differ in terms of their specific implementations of evaluating source quality based on trustworthiness and result

importance, the difference in their performance can be attributed to their importance based measures. The mediator systems are represented by the specific implementation of importance based measure. Scoring function used by importance-based source selection models uses a weighted combination of query-based relevance measure and trustworthiness measure, sourcerank, as defined in Equation 5.4. Though any query-based relevance measure can be used, CORI was used as a relevance-based measure for the experiments because of its effectiveness [16]. Importance-based source selection models are represented by the specific implementation of importance based measure and the corresponding α value i.e. the weight assigned to relevance based measure. For example, a source selection model employing TSR as an importance based measure and giving 0.9 weightage to CORI, is represented as TSR(0.9). Next section provides details for computing trustworthiness measure for these mediator systems.

Mediator System Employing Undifferentiated SourceRank, USR

The generic case of a deep-web environment is when no topic-specific information is available to differentiate between sources S, sampling queries Q_S and user queries Q_U. For such scenario, a single undifferentiated sourcerank vector is created for all sources S. *Top-5* results returned for partial sampling queries, Q_S, are used for agreement computation. These partial sampling queries were genrated by removing query terms with 0.5 probability. As the sampling queries were a mix from different representative topic-classes, an undifferentiated agreement graph AG_C was computed for the set of topic-classes C. Performing a random walk on this undifferentiated agreement graph produced an undifferentiated ranking USR of all sources S.

This undifferentiated sourcerank USR is used as part of scoring function for ranking sources for user queries Q_U posed to this generic deep-web environment.

Mediator System Employing Topic-Sensitive SourceRank, TSR

The second deep-web scenario considered is similar to the generic case except topic-specific classification is assumed to be available for sampling queries i.e. along with Q_S additional information about topic-specific sampling queries Q_{S_i} where $i \in C$ is available. *Top-5* results returned in response to partial topic-specific sampling queries are

used for topic-specific agreement computation, AG_i for each topic-class $i \in C$. The partial sampling queries were genrated by removing query terms with 0.5 probability.

Topic-specific sourceranks TSR_i, are created by performing a random walk on the topic-specific agreement graphs AG_i.

TSR approach also requires topic descriptions for identifying query-topic. For creating topic descriptions for each representative topic, complete topic-specific sampling queries, Q_{S_i} where $i \in C$, were used. Top-10 results returned by every source in response to topic-specific sampling queries were used for creating topic-specific descriptions.

Mediator System Employing Oracular Source Selection, DSR

Complete relaxation of the generic case is one where topic-specific classification is available for sources S, sampling queries Q_S and user queries Q_U. DSR assumes that a perfect classification of sources and queries are available. DSR is provided with the manually determined domain information of the sources and the test queries. A mediator system, DSR for such an environment would a combination of vertical systems, one for each of the topics. $DSR = \bigcup_{i \in C} DSR_i$

DSR represents an ideal scenario, and its performance provides an upper bound on the optimum precision that can achieved by combining relevance measure with source trustworthiness.

Each of these vertical systems is based on the availability of an oracular information that a deep-web environment DW_c exists for each topic-class c. For collecting deep-web sources for each of these deep-web environments, approach as described in section 6.1 was followed but the queries used belonged to the same topic-class for which the vertical system was being created eg. while creating deep-web environment DW_{book} for book topic-class, book titles were used as queries. During this process any sourceId which did not belong to the multi-topic deep-web environment was skipped so that sources of DW_i are a subset of S. A total of 276, 556, 572 and 281 sources were collected for DW_{camera}, DW_{book}, DW_{movie} and DW_{music} deep-web environment respectively.

SourceRanks for each vertical system DSR_c for topic-class c were created using the topic-specific sampling queries Q_{S_c}

As DSR assumes that user queries have already been classified into their respective topic-classes, during testing Q_{U_c} for topic-class c were posed to vertical system DSR_c for topic-class c.

The overall effectiveness of DSR was calculated as the sum of the effectiveness of the individual vertical systems DSR_c for each topic-class $c \in C$.

Query Similarity Based Measures

CORI

CORI is a query-based relevance measure. Source statistics for CORI were collected using highest document frequency terms from the sample crawl data. 800 high-tuple frequency terms were used as queries and *top-10* results for each query were used to create resource descriptions for CORI. Parameters found optimal by Callan *et al*. [10] were used for selecting sources based on CORI.

Google Base

TSR was compared with Google Product Search results. Two-versions of Google Base were used. [1] Gbase on dataset restricted to search only on the crawled sources, and stand alone Gbase in which Google Base search with no restriction i.e. considers all sources in Google Base.

6.5 Result Merging and Ranking

Using the source selection strategies, *top-k* sources were selected for every test query and Google Base was made to query only on these *top-k* sources. Three different values of k - *top-10* sources, *top-5%* and *top-10%* sources were used for the experiments and *k=10* was found to produce best precision and precision decreased as value of k was increased. Google Base's tuple ranking was used for ranking the resulting tuples and return *top-5* tuples in response to test queries. After ranking the tuples, the methods can be directly compared with each other.

6.6 Relevance Evaluation

Test queries defined above were used for assessing the relevance. The queries were issued to *top-k* sources selected by different source selection methods. The *top-5* results returned were manually classified as relevant or irrelevant. The classification was rule based. For example, if the test query is "Pirates Caribbean Chest" and the original movie

[1] Google Product Search implements a search on Google Base, and provides API based access as well. Though the exact searching method of Google Base in unknown, we assume that Google Base predominantly fetch results based on query similarity based on the examination of Google Base results.

name is “Pirates of Caribbean and Dead Man’s Chest“ then if the result entity refers to the movie “Pirates of Caribbean and Dead Man’s Chest“ (DVD, Blue-Ray etc.) then the result is classified as relevant and otherwise irrelevant. To avoid author bias, results from different source selection methods were merged in a single file so that the evaluator does not know which method each result came from while he does the classification.

Chapter 7

RESULTS

TSR was compared with the baselines described earlier. Instead of using stand-alone TSR, TSR was combined with query similarity based CORI measure. Experiments were conducted with different values of weighted combination of CORI and TSR, and it was found that $TSR \times 0.1 + CORI \times 0.9$ gives best precision. For rest of this section this combination is denoted as $TSR(0.9)$. Note that the the higher weightage of CORI compared to TSR is to compensate for the fact that TSR scores have much higher dispersion compared to CORI scores, and not an indication of relative importance of these measures.

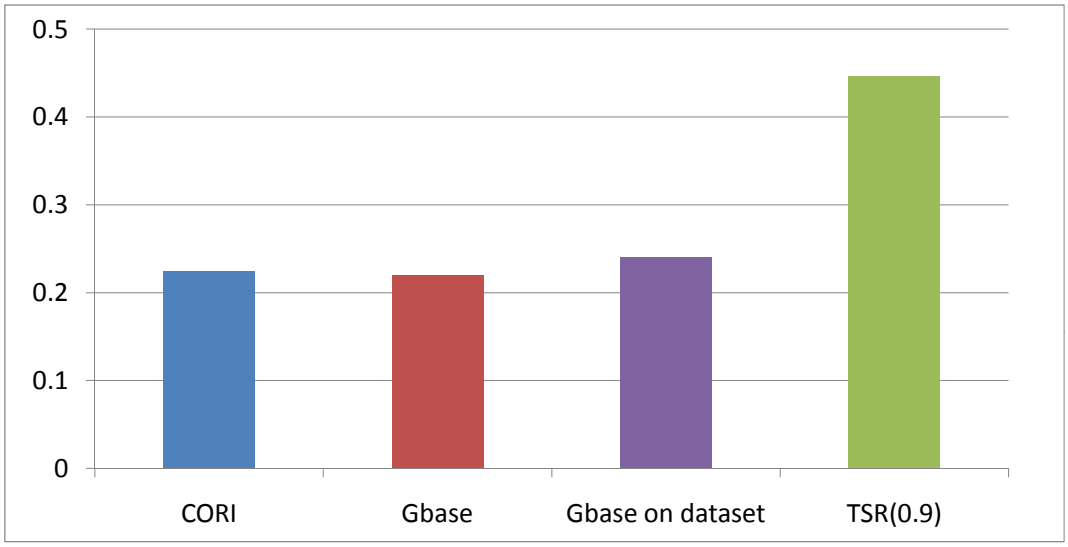

Figure 7.1: Comparison of TSR and Relevance-Based Source Selection Models

7.1 Comparison with Query Similarity Based Source Selection

The first set of experiments compare precision of TSR(0.9) with query similarity based measures i.e. CORI and Google Base discussed above. The results are illustrated in Figure 7. Note that the improvement in precision for TSR is significant as the precision improves approximately 85% over all competitors, including Google Base. This considerable improvement in precision is not surprising in the light of prior research on agreement based source selection with query based measures [6].

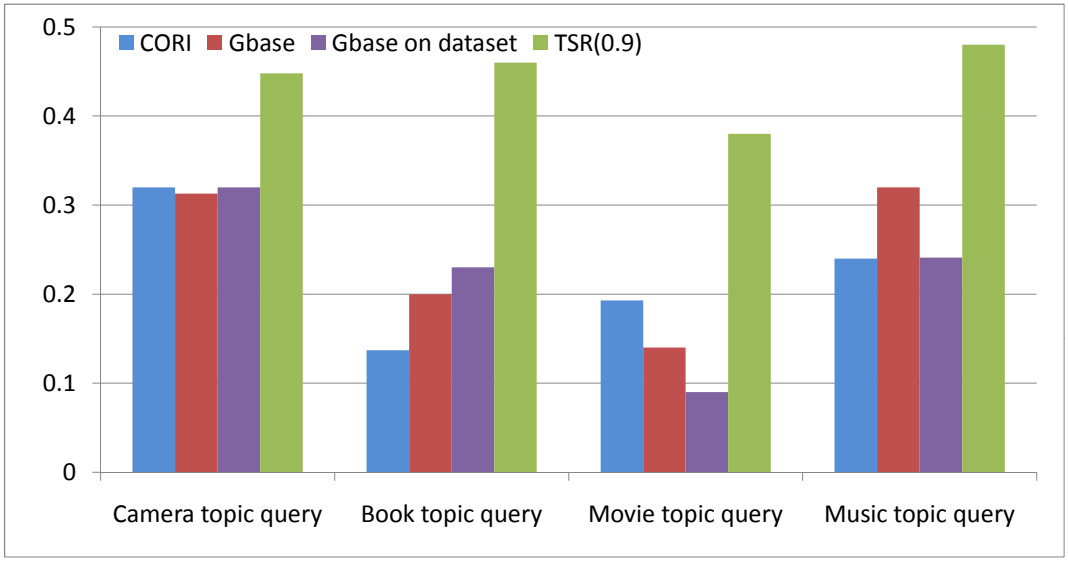

Figure 7.2: Topic-Class Based Comparison of TSR and Relevance-Based Source Selection Models

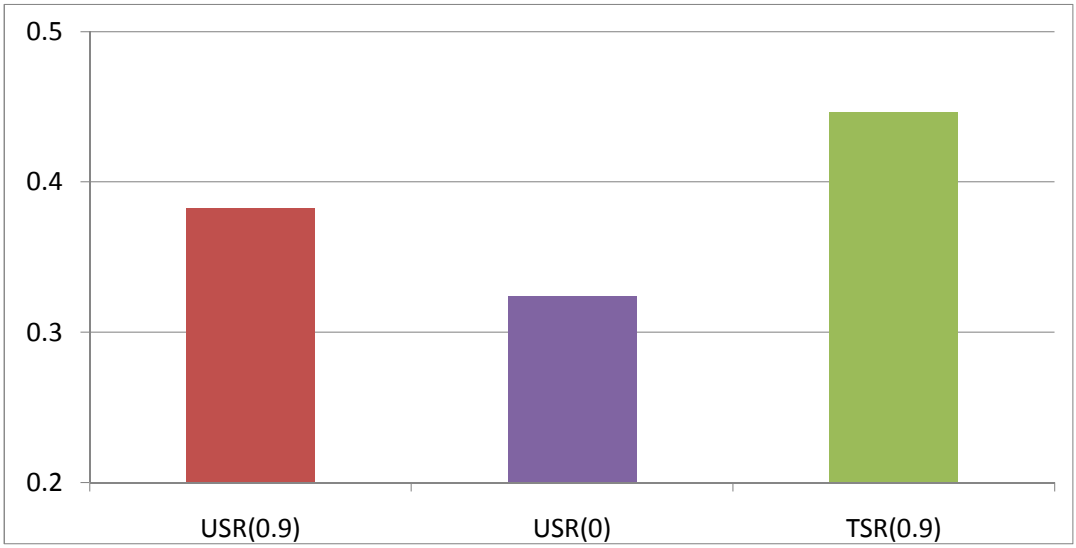

Figure 7.3: Comparison of TSR and Agreement-Based Source Selection Models

A per topic-class analysis of test queries, Figure 7, reveals that TSR(0.9) significantly out-performs the relevance-based source selection models for all topic-classes. As a note on the seemingly low precision values, these are mean relevance of the top-5 results. Many of the queries used have less than five possible relevant answers (e.g. a book title query may have only paperback and hard cover for the book as relevant answers). But since the *top-5* results always are counted, the mean precision is bound to be low. For example, if a method returns one relevant answer on in *top-5* for all

Figure 7.4: Topic-Class Based Comparison of TSR and Agreement-Based Source Selection Models

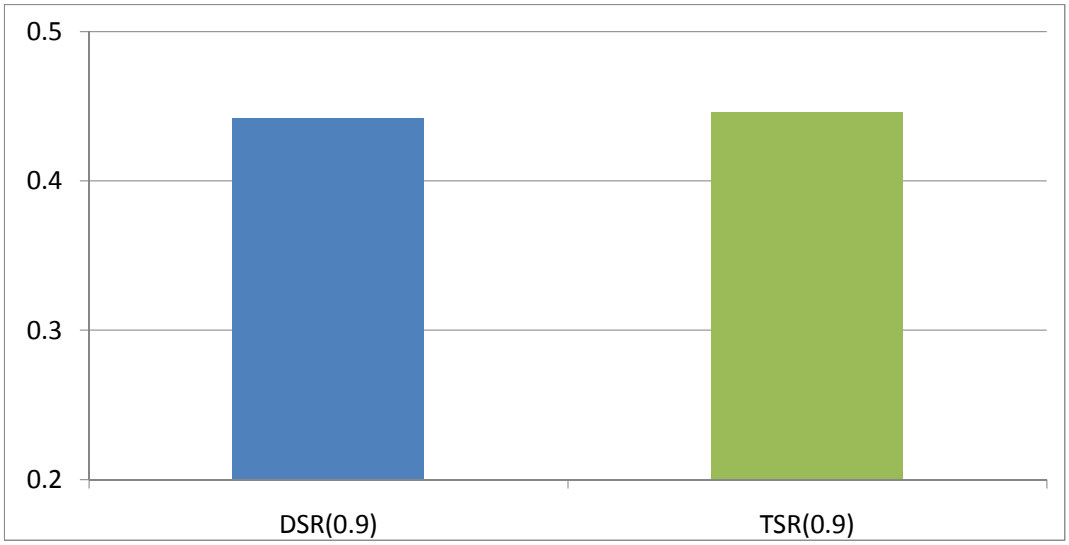

Figure 7.5: Comparison of TSR and Oracular Agreement-Based Source Selection Model

queries, the *top-5* precision value will be only 20%. Better values are obtained since some queries have more than one relevant results in *top-5* (e.g. Blu-Ray and DVD of a movie).

7.2 Comparison with Agreement Based Source Selection

TSR(0.9) is compared with the linear combination of USR and CORI.

$USR \times 0.1 + CORI \times 0.9$ was used for these comparisons. Linear combination of USR with a query specific relevance is a highly intuitive way of extending a static SourceRank multi-domain deep web search. Note that the comparison of TSR and USR is isomorphic

Figure 7.6: Topic-Class Based Comparison of TSR and Oracular Agreement-Based Source Selection Model

to the comparison of topic-sensitive PageRank [12], and PageRank [8] for the surface web.

The aggregated results across the domains are illustrated in Figure 7. TSR(0.9) precision exceeds USR(0.9) by 18% and USR(0) by 40%. Since the difference are small, the statistical significance of these results was evaluated. Sufficient number of queries were used to guarantee that TSR(0.9) out-performs both USR(0.9) and USR(0) (i.e. stand alone USR, not combining with CORI) with confidence levels of 0.95 or more.

Figure 7 provides per topic results. For three out of four topic-classes (Camera, Movies, and Music), TSR(0.9) out-performs USR(0.9) and USR(0) with confidence levels 0.95 or more. For books no statistical significant difference was found between USR(0.9) and TSR(0.9). This may be attributed to the fact that the source set was dominated by large number of good quality book sources, biasing the ranking towards book domain. Further, analysis revealed that there are many multi-domain sources providing good quality results for books, movies and music domains (e.g. Amazon, eBay). These versatile sources occupy top positions in USR as well as USR(0.9) for these three domains. Consequently the domain independent USR performs comparable to domain specific USR(0.9) for these three domains: music, movies and books.

7.3 Comparison with Oracular Agreement Based Source selection

In the next set of experiments, TSR was compared with oracular source selection, DSR described above in earlier. TSR(0.9) was compared with DSR(0.9) (i.e. linear combination $0.1 \times DSR + 0.9 \times CORI$). As shown in Figures 7 and 7, TSR(0.9) is able to match DSR(0.9) performance for the test queries. The aggregate results across the domains is shown in Figure 7 and domain-wise result is shown in Figure 7. Result shows that the TSR precisions are quite comparable with that of DSR. This implies that TSR is highly effective in categorizing sources and queries, almost matching with oracular DSR.

Chapter 8

CONCLUSION

In this work, an attempt was made to perform multi-domain source selection sensitive to trustworthiness and importance for the deep web. Although SourceRank, which considers source trustworthiness and importance in assessing source quality, is effective in single-topic environments, the need for extending it to multi-topic deep-web environments was discussed. To help understand the problem of deep-web environments, a way of representing a deep-web environment was formulated. Essential properties of an importance measure for a multi-topic deep-web environment were also defined. Based on the two most popular surface-web's linked based techniques, different ways were explored for extending sourcerank to multi-topic deep web environment. Topic-sensitive SourceRank (TSR) was introduced as an efficient and effective technique for evaluating source importance in a multi-topic deep web environment. TSR source selection was combined with a Naïve Bayes Classifier for queries to build the final multi-domain deep web search system. Experiments on a more than thousand sources spanning across multiple topics shows that a TSR-based source selection is highly effective in extending SourceRank for multi-domain deep web search. TSR is able to significantly out-perform query similarity based retrieval selection models including Google Product Search by around 85% in precision. Comparison with other baseline agreement-based source selection models showed that using TSR results in statistically significant precision improvements over baseline methods; including a domain oblivious SourceRank combined with query similarity. Comparison with oracular DSR approach reveals effectiveness of TSR for domain-wise query and source classification and subsequent source selection.

REFERENCES

[1] Ny times. http://www.nytimes.com/.

[2] Open directory project. http://www.dmoz.org.

[3] Pbase. http://www.pbase.com/.

[4] Wikipedia. http://www.wikipedia.org/.

[5] Yahoo directory. http://dir.yahoo.com/.

[6] R. Balakrishnan and S. Kambhampati. Sourcerank:relevance and trust assessment for deep web sources based on inter-source agreement. *WWW*, pages 237–246, 2011.

[7] M Bergman. The deep web: Surfacing hidden value, 2000.

[8] S. Brin and L. Page. The anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30, April 1998.

[9] J. Callan and M. Connell. Query-based sampling of text databases. *ACM TRANSACTIONS ON INFORMATION SYSTEMS*, 19, 1999.

[10] J. Callan, Z. Lu, and B. Croft. Searching distributed collections with inference networks. *SIGIR '95 Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995.

[11] L. Gravano, H. Garcia-Molina, and Tomasic A. The effectiveness of gioss for the text database discovery problem. *SIGMOD '94 Proceedings of the 1994 ACM SIGMOD international conference on Management of data*, 23, June 1994.

[12] T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15, 2003.

[13] Madhavan J., Ko D., Kot L., Ganapathy V., Rasmussen A., and Halevy A. Google's deep web crawl. *Proceedings of the VLDB Endowment*, 1, August 2008.

[14] J. Klienberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, September 1999.

[15] Z. Nie and S. Kambhampati. A frequency based approach for mining coverage statistics in data integration. *Proceedings of ICDE*, 2004.

[16] A. Powell and J. Fench. Comparing the performance of collection selection algorithms. *ACM Transactions on Information Systems (TOIS)*, 21, October 2003.

[17] M Shokouhi and J. Zobel. Federated text retrieval from uncooperative and overlapped collections. *SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007.

[18] L. Si and Callan J. Relevant document distribution estimation method for resource selection. *SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.

[19] A. Wright. Searching deep web. *Communications of the ACM*, 51:14–15, November 2008.